UK Research and Innovation

DAFNI

Use case for Water Systems Leakage (WSL) for the Data Infrastructure for National Infrastructure project (DINI)

Ruoqing Yin, Haonan Xu, Jiaqian Wei, Liz Varga

Contents

1. Use case Report

1.1 Background and Context

With almost 20% of water resources wasted in leakage, the aim of this project is to use data to enhance leakage detection in water distribution systems, to identify barriers to data sharing, and to propose solutions to facilitate co-operation between stakeholders.

The project operates within the water utility sector and involves different stakeholder groups:

- Water utilities: The main data provider for operations, maintenance and sensor data.
- Regulators: Responsible for monitoring performance standards and encouraging leakage reduction, e.g. Ofwat.
- Local authorities and councils: Key beneficiaries of actionable insights into infrastructure maintenance and resource allocation, policy makers.
- Research organisations: Collaborators providing expertise in artificial intelligence, machine learning and hydraulic modelling for advanced leakage detection methods, whose research is also affected by data barriers.

This project investigates how to simultaneously overcome data sharing constraints and utilise operational data to optimise leakage detection while allowing the above users to benefit through data sharing.

1.2 Description of Activities

Build Data Ontology:

- Collect operational data, company reports, etc. from public databases, water supply companies. Use this data to establish data standards and ontology.

Identify Data Barriers:

- Browse relevant academic literature and company reports, identify relevant data barriers in the literature and establish possible solutions.

Stakeholder Engagement:

- Conduct workshops and interviews with utilities, academics, and technicians to understand their data-sharing practices and concerns and validate whether data barriers hold true for different stakeholders and the feasibility of solutions.

- Engage with regulators to align project goals with policy frameworks.

- Collaborate with academic institutions to develop AI-based tools for leak detection.

Data Sharing Analysis:

- Map existing data sharing protocols and identify gaps, such as legal restrictions (GDPR) and barriers such as commercial sensitivity.

1.3 Benefits of Data Sharing

Data sharing will provide benefits to relevant stakeholders across the water sector, such as water companies, researchers, government, regulators, etc.

- Improved water leakage detection:
  Real-time sensor data enables early and proactive detection of leaks rather than relying on user reports, reducing water loss and maintenance costs.
- Enhanced research collaboration:
  Sharing data fosters innovation and enables researchers to validate models using real-world scenarios.
- Policy support:
  Transparent access to leakage data informs evidence-based decision-making and contributes to sustainable development goals such as reducing carbon emissions.
- Future Trends:
  Open data and model training is an increasingly important trend, and promoting standardised data sharing will provide companies with future competitiveness
- Operational Efficiency:
  Utility companies can benefit from optimised leakage prioritisation and cost-effective maintenance scheduling.
- Scalability:
  Standardisation of datasets allows replication of solutions across regions and systems and promotes the potential for mutual aid and exchange.
- Increased public trust:
  Open data sharing increases accountability and demonstrates utilities' efforts to address water loss.

1.4 Barriers for Data Sharing

The project identified the following main barriers to effective data sharing in the water sector and uses the DAFNI classification system repeated in Table 1.

Table 1: DAFNI classification system for barriers to data sharing

	Barrier	Description
1.1	Personal Sensitivity	The data has additional restrictions due to personal sensitivity
1.2	Commercial Sensitive	The data has additional restrictions due commercial sensitivity
1.3	Legal Ownership	It isn't clear who owns the data. The ownership of the data restricts access
2.1	Organisational Security	Is the data restricted due to organizational security concerns
2.2	National Security	Is the data restricted due to national security concerns?
3.1	Cost / Economics	There are financial costs to access the data. The resources needed to make the data useable for your purpose outweigh the benefits of using this data
3.2	Licensing	The license assigned to the data means that it can't be used for your purpose. There is no license assigned so that the data is riskier to use for your purpose
3.3	Contractual issues (DSA's)	The Data Sharing agreement is not acceptable to your organization. It takes too long for the DSA to be agreed for it to be used for your purpose
4.1	Discoverability	It is difficult to find or access available data?
4.2	Cultural Barriers	Is there indifference or resistance obstructing data sharing?

4.3	Ethics	Are there ethical consequences of the data being shared – e.g. unfair treatment or adverse consequences to particular stakeholder groups.
4.4	Lack of appropriate skills	The skills are not available at the data source or by the data user to access and use the data ?
	Barrier	Description
5.1	Reliability	Is the data accurate and relevant? Is it out of date? Does it come with quality metrics?
5.2	Interoperability/ standards	The data is in file formats that are difficult to share (proprietary) or there are no standards applied to the data
5.3	Metadata	The data is not well described and so it is difficult to find and use.
5.4	Quality measures	The data has not undergone any quality control. The quality control measures are not documented
5.5	Lack of suitable computing support	Are the computing systems insufficient to support the effective sharing and reuse of data?
5.6	Contextual information	There is not enough information associated with the data resource to enable someone outside the dcata creator/generator group to use it
5.7	Coverage	Is the data accurate, but does not give the coverage required? This can be coverage in time, or in geographic coverage or in resolution?

DINI WSL barriers are:

- Legal and commercial security restrictions (1.1, 1.2, 1.3 and 2.1):
  GDPR and commercial sensitivities limit access to real business data, thus limiting collaboration and transparency.

- Limited public datasets (4.1):
  Data access for the general public and academic researchers is often extremely limited, affecting the validation of methods.

- Data quality issues due to cost constraints or technical limitations (3.1, 5.4 and 5.5): Inadequate sensor coverage and configuration issues will have an impact on data quality. In addition, sensor data is often incomplete, noisy etc. due to transmission and equipment failure issues, reducing its usefulness for advanced analysis and modelling, again quality issues are a reason why a water company would be reluctant to share it externally.

- Lack of standardisation (5.2): Inconsistent units, sampling rates and terminology hinder interoperability and comparative studies and prevent data sharing between water companies.

- Cultural resistance (4.2): Organisations are reluctant to share data for fear of competitive advantage or data misuse.

- Technical barriers (5.3, 5.5 and 5.6): Challenges in ensuring interoperability between different data formats and systems. Metadata and contextual information is often insufficient, making datasets more difficult to interpret, and unlabelled datasets interfere with supervised training.

Recommendations to overcome these barriers include :

- Legal framework: Develop data sharing agreements that protect commercial interests while allowing researchers access.
- Standardisation initiatives: Adopt data standards such as the IWA standard to standardise units, terminology and sampling protocols.
- Anonymisation techniques: Use anonymised datasets to balance privacy concerns with data availability.
- Capacity Building: Train stakeholders on the benefits and methods of data sharing and eliminate cultural resistance.

1.5 Sources of data

The project identified the diverse sources of data on WSL as shown in Table 2.

Table 2: Sources of WSL data

Data Source	Data Description	Purpose	Technical Details	Data restrictions and Licence	Barrier	Stakeholder
District metered area link	DMA_ID: Unique identifier for each District Metered Area (DMA). CENTROID_X: X-coordinate of the geographical center of the DMA. CENTROID_Y: Y-coordinate of the geographical center of the DMA. Geometry: DMA boundaries.	The purpose of accessing DMA locations is to facilitate precise geographical identification and analysis of water leakage patterns within each District Metered Area (DMA). By leveraging spatial coordinates and boundary data, utilities can efficiently localize and address leaks, optimize resource allocation, and enhance the overall resilience of the water distribution system	Polygon geometry, typically formatted using spatial data standards such as GeoJSON, WKT (Well-Known Text), or Shapefile to represent the DMA boundaries	No License Provided Request permission to use	4.4	Water Companies, Regulators, Researchers, Data Aggregators
Water Consumption Data Link	Data related to household and non-household water usage, population, meter readings, and aggregated water consumption across DMAs.	To analyze water usage patterns, detect anomalies such as leaks, optimize resource allocation, and improve water distribution efficiency.	Data includes identifiers, consumption metrics, and temporal information. Aligned with geographic and temporal scales, using formats like integers, floats, and timestamps.	- Open Data Commons License for some datasets. - No license provided for others; permission required. - Data accuracy depends on aggregation methods and years.	5.1, 5.2, 5.3, 1.1	Water Companies, Regulators, Researchers, Data Users, Data Aggregators

Geological Data Link	Information on soil texture, pH corrosivity, permeability, and density. These attributes describe soil properties relevant for understanding environmental factors influencing water leakage in pipelines.	To identify areas prone to leaks due to soil conditions, enabling precise maintenance strategies, reducing water loss, and optimizing water distribution networks.	Data includes soil classifications (e.g., sandy, clay) and numeric values for pH, permeability (cm/hr), and density (g/cm^3). The data is linked to geographic coordinates and spatial resolutions at defined depths.	Open Government Licence. Accessible for use with attribution.	1.2, 4.4	Researchers, Water Companies, Regulators
Customer Complaint Data Link	Data on customer complaints, including complaint IDs, date received/sent, narrative description, postcodes, complaint categories, timely response indicators, and handling company names.	To identify specific areas and patterns of water leakage through customer feedback, enabling targeted investigations, resource allocation, and faster resolution of water-related issues. Supports customer satisfaction and operational efficiency.	Includes alphanumeric complaint IDs, dates (YYYY-MM-DD), text narratives, complaint categories, and binary fields (yes/no). Data aligns with privacy standards (e.g., personal data redaction).	Creative Commons Zero v1.0 Universal Licence. Compliance with privacy standards required for narrative text fields.	1.1, 1.2, 4.3, 5.2, 5.3	Water Companies, Regulators, Policy Makers, Researchers
Sensor location Link	Geographical sensor data including unique identifiers (SENSOR_ID) and X, Y coordinates for sensor placement.	To map sensor locations accurately within the water distribution network, enabling monitoring of areas affected by leaks, early leak detection, and optimized maintenance strategies.	Includes unique string identifiers (SENSOR_ID) and decimal coordinates (X, Y) for geographical mapping.	License for the dataset is unspecified.	5.1, 4.1, 2.2	Water Companies, Data Infrastructure Providers, Researchers, Policy Makers
Measured data Link	Geographical sensor data including unique identifiers (SENSOR_ID) and X, Y coordinates for sensor placement.	To map sensor locations accurately within the water distribution network, enabling monitoring of areas affected by leaks, early leak detection, and optimized maintenance strategies.	Includes unique string identifiers (SENSOR_ID) and decimal coordinates (X, Y) for geographical mapping.	License for the dataset is unspecified.	5.1, 4.1, 2.2	Water Companies, Data Infrastructure Providers, Researchers, Policy Makers

Sensor Asset Data Link	Data on the lifecycle and operational status of sensors, including age, type, manufacturer, installation year, operational status, and model classification.	To monitor the lifecycle, reliability, and operational status of sensors for proactive maintenance, optimized deployment, and improved efficiency of water monitoring systems.	Includes age (years), type (e.g., flow, pressure), installation year (YYYY), status (active, inactive, maintenance), and manufacturer/model details. Standardized fields ensure uniformity and accuracy across records.	Open Government Licence v3.0 for general data. MIT License for API usage.	5.1, 4.1, 5.7, 5.3, 4.4	Water Companies, Data Infrastructure Providers, Researchers, Manufacturers
Pipe Location and Asset Data Link	Data about pipeline locations, physical attributes, and operational status. Includes pipe ID, geometry, length, material, diameter, operational status, year of installation, number of breaks, and flow direction.	To analyze the physical and operational status of pipelines for identifying vulnerabilities, predicting failures, and improving maintenance strategies. Supports infrastructure management to reduce leakages and enhance service reliability.	Data includes spatial geometry (line objects), numerical attributes (length in meters, diameter in cm), categorical fields (material, status, flow direction), and installation year (YYYY format).	Licensing information partially available; access for some datasets may require permissions.	5.1, 4.1	Water Companies, Regulators, Infrastructure Managers, Researchers
Repair Data Link	Data related to repair activities, including repair ID, status, location, timelines (start and end dates), repair type, last maintenance date, and fault type.	To track and analyze repair activities, enabling targeted maintenance, reducing downtime, and enhancing the resilience and reliability of the water distribution network.	Includes unique identifiers, dates/times (ISO 8601 format), geographical coordinates (latitude/longitude), and standardized repair/fault types. Data aligns with maintenance and fault classification standards.	Creative Commons BY-SA 4.0 License. Some data access restrictions remain unspecified.	5.1, 4.1	Water Companies, Maintenance Teams, Infrastructure Managers, Researchers

1.6 Results Data

Information on the data generated as a result of the research undertaken in the project are provided below.

1. Data Description

- **Data**
  - **Data Types**:

        - Leakage data (quantitative, time-series).
        - Anomalies in smart meter data.
        - Leakage calculations and AMP8 target progress.
        - Outputs from analysis tools (e.g., Minimum Night Flow and Mass Balance methods).
    - **File Formats**:
        - CSV/Excel for raw data.
        - PDF/Word for reports and analyses.
        - Spatial data (e.g., DMA polygons) in GIS-compatible formats.
    - **Volume and Frequency**:
        - Large datasets with 15-minute to 1-hour granularity, spanning multi-year periods.
    - **Size:**
        - Varies from MBs for specific subsets (e.g., DMAs, polygons) to TBs for long-term, high-frequency raw data.
    - **Ontologies and Standards:**
        - Use domain-specific ontologies to structure and standardize data:
            - HydroOntology for water-related terms.
            - INSPIRE Data Specifications for spatial data.
        - Implement semantic tagging to improve interoperability between datasets.
- **Persistent Identifiers (PIDs)**:
    - Assign **DOIs (Digital Object Identifiers)** to datasets, particularly for public repositories, to ensure long-term discoverability.
    - Use unique identifiers for internal datasets (e.g., UUIDs for specific files or records).
- **Rich Metadata**:
    - Include detailed metadata describing:
        - **Dataset Title**: Clear, descriptive title.
        - **Description**: Purpose, methods, and context of the data.
        - **Keywords**: Relevant terms (e.g., "water leakage," "DMA polygons," "sensor data").
        - **Temporal/Spatial Coverage**: Time ranges, DMA boundaries, and geographic context.
        - **Data Creators**: Attribution for researchers and institutions involved.
    - Use metadata standards such as **Dublin Core** or **ISO 19115** for geospatial datasets.
    - Machine-Readable Metadata:
        - Provide metadata in machine-readable formats (e.g., XML, JSON-LD).
        - Link datasets to external resources using Linked Data principles.
- **Indexing**:
    - Ensure datasets are indexed in university repositories, public repositories (e.g., Zenodo), or domain-specific catalogs (e.g., UKWIR).

2. Data Collection and Processing

- **Sources**:
  - Smart meters, acoustic sensors, and satellite imagery.
  - Public datasets (e.g., NASA precipitation data).
  - Outputs from AI/ML algorithms.
- **Quality Assurance**:
  - Data cleaning and validation to address gaps and inconsistencies.
  - Automated QA processes where feasible (e.g., AI-based).

3. Data Storage and Backup

- **Storage**:
  - Sensitive data: Use university repositories or trusted platforms with strict access controls.
  - Open-access data: Publish via public repositories like Figshare or Zenodo with proper metadata and licenses. Assign DOIs (Digital Object Identifiers) to datasets, particularly for public repositories, to ensure long-term discoverability.
  - Use unique identifiers for internal datasets (e.g., UUIDs for specific files or records).
  - Large-scale industry data: Store in cloud-based systems with encryption or centralized industry repositories for domain-specific use.
- **Backup Schedule**:
  - Weekly backups for all datasets.
  - Hourly backups during active project phases.

4. Data Sharing and Access

- **Platforms**:
  - Open-access repositories for non-sensitive data.
  - Controlled platforms for sensitive and proprietary data (e.g., STREAM Secure Research Environment).
- **Stakeholders**:
  - Water companies, researchers, regulators (e.g., Ofwat, EA).
- **Access Levels**:
  - Tiered permissions (e.g., public, restricted, confidential).
  - Anonymization for sensitive data (e.g., customer-level leakage data).
  - Explain how data sharing agreements and access levels align with FAIR principles (open or controlled access)
  - **Open Access**: Non-sensitive, anonymized datasets.
  - **Controlled Access**: Sensitive datasets requiring approval via secure platforms.

  - Use APIs or data-sharing portals for easy retrieval of accessible data.
  - **Long-term Access:**
    - Store data in repositories with guarantees of long-term accessibility (e.g., university repositories, Zenodo).
- **Restrictions on Data Sets**
  - **Transferred Requirements from Data Suppliers:**
    - Data sharing agreements often include non-disclosure agreements (NDAs) to protect proprietary and sensitive information provided by water companies and stakeholders. These may restrict how data can be shared or reused.
    - Some data sets, such as raw sensor data and pipeline infrastructure details, may be subject to ownership rights by the water companies, requiring explicit permissions for use beyond the original project scope.
    - Industry-specific data, like DMA boundaries or pipeline conditions, may have regional variations in regulations that require compliance with local policies before sharing.
  - **Personal Sensitivity:**
    - Household-level water usage data collected via smart meters or acoustic sensors may reveal activity patterns, raising privacy concerns. Anonymization and data minimization techniques must be applied to protect individuals' privacy.
    - Metadata from smart meters could inadvertently expose behavioral patterns, such as appliance usage or occupancy levels, necessitating careful data aggregation to ensure no personally identifiable information (PII) is accessible.
    - Advanced technologies like hydrophones and accelerometers may exacerbate privacy risks by capturing unintended information, requiring the use of block charge systems or similar measures to anonymize data.
- **Who is the Data Going to Be Shared With?**
  - Water Companies: The data will be shared with water companies for operational improvements, such as leakage detection, infrastructure planning, and meeting regulatory compliance targets.
  - Regulators (e.g., Ofwat): Specific datasets, especially those related to leakage rates and compliance metrics, will be shared with regulators to ensure adherence to industry standards and support policy development.
  - Industry Partners: Trusted partners involved in pipeline repairs, smart metering rollout, and infrastructure testing (e.g., contractors or Meter Asset Providers) will have controlled access to relevant datasets to improve operational efficiency and drive innovation.
  - Research Collaborators: Academic and industry researchers focusing on water resource management and leakage prevention will have access to anonymized or aggregated datasets under strict data-sharing agreements.

5. Legal and Ethical Considerations

- **Ownership**:
  - Clearly define data ownership between stakeholders to avoid conflicts.
- **Privacy**:
  - Anonymize customer-related data to protect privacy.
  - Adhere to GDPR and local privacy regulations.
- **Licensing**:
  - **Apply clear licensing terms, such as:**
    - Creative Commons Licenses (CC BY) for open data.
    - Restricted-use licenses for sensitive data, specifying permissible use (e.g., for research only).
  - Include a data-sharing agreement for industry partners to clarify terms of access and liability.

6. Data Utilization and Outputs

- **Reusable**
  - Provenance:
    - Document the data lifecycle, including:
      - Source: Where the data was collected (e.g., specific sensors or DMAs).
      - Methods: How the data was processed or analyzed.
      - Quality Assurance (QA): Steps taken to validate data accuracy and integrity.
      - Version Control: Maintain clear versioning for iterative datasets.
  - Documentation:
    - Provide comprehensive documentation, including:
      - A data dictionary explaining variable names, units, and codes.
      - Step-by-step guides for accessing and using the data.
- **Deliverables**:
  - Leakage detection models and algorithms.
  - Reports on the effectiveness of mitigation strategies.
  - Visualizations (charts, GIS maps).
- **Publishing**:
  - Academic journals, conference presentations, and industry white papers.

7. Long-Term Preservation

- **Archiving**:
  - Long-term storage in national or institutional repositories (e.g., UKDA).
  - Metadata tagging for easy retrieval.

- **Retention Period**:
    - Minimum of 10 years post-project completion.

8. Responsibilities and Resources

- **Roles**:
    - Data Manager: Oversees data handling and compliance.
    - IT Support: Manages storage and backup infrastructure.
- **Training**:
    - Stakeholder training in data handling, privacy, and analysis.

9. Review and Updates

- **Frequency**:
    - Quarterly reviews of the DMP.
- **Adjustments**:
    - Update the DMP to reflect changes in regulations, technology, or project scope.

1.7 Lessons Learnt and Recommendations

Provide information on lessons learnt and recommendations identified by the project. The following questions may help complete this section:

- ***What has been learnt over the course of the project?***
    - Over the course of the project, it became evident that consistent data collection and standardization are critical for meaningful analysis. Challenges around privacy, data ownership, and infrastructure vulnerabilities highlighted the need for early agreements on sharing protocols and clear roles among stakeholders.
- ***What works well at the moment?***
    - Smart metering technology and real-time flow monitoring have proven effective in detecting and isolating leaks. Additionally, trusted research environments and collaborative platforms (e.g., STREAM) facilitate secure data sharing while maintaining confidentiality.
- ***Please provide some examples of effective data sharing.***
    - Successful implementation of anonymized leakage datasets shared with academic researchers allowed for advanced AI modeling to predict high-risk areas.
    - Collaborative efforts between water companies and contractors enabled joint access to infrastructure data, leading to faster detection and repair of leaks.
- ***What are the priorities identified for resolving any barriers?***

  - Addressing privacy and legal challenges through standardized NDAs and anonymization techniques.
  - Improving workforce training to bridge technical knowledge gaps in utilizing advanced tools and systems.
  - Aligning regional regulatory policies to streamline data governance and sharing.
- ***If there was future funding, how would this project progress?***
  - With future funding, the project would scale up meter data improvements, integrate AI-based leakage detection technologies, and enhance secure platforms for sharing sensitive infrastructure data. Additional resources would support advanced workforce training and pilot programmes for innovative leak prevention strategies.

Appendix 1 WSL Wiki

Data Source			Data Description	Purpose	Technical Details			Data restrictions and Licence	Barrier
type	name	link			data type	UNITS	Data standard Description		
DMA locations	DMA_ID	https://hub.arcgis.com/datasets/f1aeaa7ad2c947048eaf9fc06b6df0e5/explore	Unique identifier for each District Metered Area (DMA)	The purpose of accessing DMA locations is to facilitate precise geographical identification and analysis of water leakage patterns within each District Metered Area (DMA). By leveraging spatial coordinates and boundary data, utilities can efficiently localize and address leaks, optimize resource allocation, and enhance the overall resilience of the water distribution system.	Integer	/	Follows a unique alphanumeric identifier format as per the standard for DMA identification	No License Provided Request permission to use	4.4
	CENTROID_X_		X-coordinate of the geographical center of the DMA		decimal	/	Standard coordinate format for geographical data, typically using a specified coordinate system such as WGS84 or UTM		
	CENTROID_Y_		Y-coordinate of the geographical center of the DMA		decimal	/			
	Geometry	https://hub.arcgis.com/datasets/f52d516e8e4b4bd0913062c4796ea32d_0/explore?location=-41.187044%2C175.119137%2C10.00	DMA boundaries		Spatial Object	/	Polygon geometry, typically formatted using spatial data standards such as GeoJSON, WKT (Well-Known Text), or Shapefile to represent the DMA boundarie	The data has been compiled from a variety of sources and its accuracy may vary.	4.4
consumption	num_household	https://datamillnorth.org/dataset/2zlgn/yorkshire-water-leakage-dma-15-minute-data	Number of Household Properties within the DMA	For the provided data consumption fields, the purpose is to analyze water usage patterns across households and non-households within each DMA, which is crucial for identifying anomalies that may indicate leaks. Accessing this data enables effective tracking of water flow, supports targeted leakage detection, and helps optimize resource allocation within water distribution networks	Integer	/	Integer format representing the count of household properties within each DMA. Each value corresponds to a specific DMA or Area identifier like a postcode.	https://opendatacommons.org/licenses/by/1-0/	4.5, 5.1, 5.3, 5.2
	num_nonhousehold		Number of Non-Household Properties within the DMA		Integer	/			
	DMA_population	https://hub.arcgis.com/maps/3f5ec1f4ba054421ba1b1ab303d3db5c/explore?location=37.743222%2C-122.417180%2C12.49	Population within the DMA		Integer	/	Population values are recorded for specific years (e.g., 2016, 2000) and should align with the designated DMA or spatial area identifier	No License Provided Request permission to use	5.1, 5.2, 5.3, 4.4
	meter_id	https://datamillnorth.org/dataset/2jqzm/customer-meter-data	Unique identifier for each meter		Integer	/	Unique identifier for each water meter, corresponding to the geographical identifier (e.g., POSTCODE) and associated with a specific property type (e.g., House, Domestic Properties)	https://opendatacommons.org/licenses/by/1-0/	5.1, 5.2, 5.3, 1.1
	reading_start_date		Date when the water meter reading began		Datetime	/	Start/end date and time of water meter readings, linked to the geographical identifier (e.g., POSTCODE) and relevant property type, in a standard datetime format DD/MM/YYYY HH		
	reading_end_date		Date when the water meter reading ended		Datetime	/			
	reading_start_reading		Initial reading of water usage at the start date		Float	m3 (1000 litres)	Initial/end meter reading of water usage at the start/end date, corresponding to the geographical identifier and specific property type, in units of water consumption		
	reading_end_reading		Final reading of water usage at the end date		Float	m3 (1000 litres)			
	totol_consumption_by_DMA	https://www.streamwaterdata.co.uk/datasets/f2cdc1248fcf4fd289ac1d3f25e75b3b_0/explore	Total water consumption within each DMA for a specific year, based on aggregated meter readings		Float	m3 (1000 litres)	Recorded in cubic meters (m^3) and associated with each DMA . Consumption data is annual, corresponding to the year specified, alignment with both the geographical identifier (DMA) and the temporal scale (Year)	CC BY 4.0 License	not identified
	num_meter_by_DMA		Number of active water meters within each DMA for a given year		Float	m3 (1000 litres)	Integer format, representing the count of meters per DMA, linked to the geographical identifier and the specific temporal scale		

geological	soil_texture	https://mapapps2.bgs.ac.uk/ukso/home.html	Texture classification of the soil (e.g., sandy, clay)	For geological data, accessing information like soil texture, corrosivity, permeability, and density is essential to understand the environmental factors that influence water leakage in pipelines. This data helps identify areas prone to leaks due to soil conditions, enabling more precise maintenance strategies and reducing water loss in the distribution network.	String	/	Categorical classification of soil texture (e.g., sandy, clay), associated with specific geometry coordinates at a defined depth and spatial resolution	https://www.bgs.ac.uk/bgs-intellectual-property-rights/open-government-licence/	1.2, 4.4
	soil_corrosivity (ph)		pH level indicating the corrosiveness of the soil		Float	pH (standard pH scale)	Numeric pH value representing soil corrosivity, linked to geometry coordinates at a specified depth and resolution		
	soil_permeability		Rate at which water can move through the soil		Float	cm/hr	Numeric value indicating the rate of water movement through the soil, associated with geometry coordinates and measured at a given depth and spatial resolution		
	soil_density		Density of the soil in the DMA area		Float	g/cm^3	Numeric value representing soil density, tied to specific geometry coordinates in the DMA and measured at a designated depth and resolution		
customer complaint	complaint_id	https://www.ofwat.gov.uk/regulated-companies/company-obligations/performance/companies-performance-2014-15/customers/	Unique identifier for each customer complaint	The purpose of accessing customer complaint data is to identify specific areas and patterns of water leakage based on customer feedback, enabling targeted investigations and quicker resolutions. This data supports efficient resource allocation and improves customer satisfaction by addressing leakage issues in a timely and structured manner.	Integer	/	Unique alphanumeric identifier for each complaint	Creative Commons Zero v1.0 Universal	1.1, 1.2, 4.3, 5.3 5.2
	date_received	https://www.ofwat.gov.uk/wp-content/uploads/2015/10/rpt_com201503sim.pdf	Date the complaint was received		Datetime	/	Date when the complaint was received, recorded in the standard date format (e.g., YYYY-MM-DD)		
	date_send_to_company	https://www.ofwat.gov.uk/wp-content/uploads/2018/07/Ofwat-Non-Household-Customer-Insight-Survey-2018-Wave-2-Final-Results.pdf	Date the complaint was sent to the relevant company for action		Datetime	/	Date when the complaint was forwarded to the relevant company, following a standard date format (e.g., YYYY-MM-DD)		
	narrative		Free test. Description of the customer complaint		String	/	Free-text field allowing detailed description of the customer complaint. Text should be sanitized for privacy (e.g., redaction of personal information) in compliance with data protection standards		
	postcode		Postcode of the area where the complaint originated		String	/	Postal code of the area where the complaint originated		
	complaint_category	https://www.affinitywater.co.uk/reportaproblem	The type of issue reported by the customer, categorized for easier identification and handling (e.g., 'Water leaks,' 'No water,' 'Water pressure issues')		String	/	Categorical field representing predefined complaint types, using standardized labels to ensure consistency across records. Each category corresponds to a specific water-related issue, enabling structured data analysis and facilitating efficient response handling		
	timely_response		Whether the company gave a timely response		String	/	yes/no		
	company	https://huggingface.co/datasets/CFPB/consumer-finance-complaints	Name of the company handling the complaint		String	/	Name of the company handling the complaint		
Sensor Data									

Sensor locations	SENSOR_ID		Unique identifier for each sensor	The purpose of accessing sensor location data is to accurately map the placement of sensors within the water distribution network, facilitating the identification of areas affected by leaks. This data enables precise monitoring, quicker detection of anomalies, and efficient deployment of maintenance resources to minimize water loss.	String	/	Unique identifier for each sensor	The license for this dataset is unspecified	5.1, 4.1, 2.2
	X		X-coordinate of the sensor's geographical location		decimal	/	coordinate of the sensor’s geographical location		
	Y	https://data.cityofchicago.org/Environment-Sustainable-Development/Array-of-Things-Locations/6rq2-yx28/about_data	Y-coordinate of the sensor's geographical location		decimal	/			
measured data	date/time_measured		Date and time when the data was recorded by the sensor	The purpose of accessing measured sensor data is to monitor real-time water pressure, flow, and temperature, which are crucial for detecting deviations that may indicate leaks or pipeline issues. This data allows for detailed temporal analysis, enabling rapid identification and resolution of anomalies to improve the efficiency and reliability of the water distribution system.	Datetime	/	Date and time stamp in ISO 8601 format (YYYY-MM-DD HH:MM to capture the exact moment of data recording for temporal analysis	https://www.opendatacommons.org/licenses/by/1.0/index.html	5.1, 5.4, 5.2, 5.3
	pressure_data		Water pressure data recorded by the sensor		Float	usually Pa	Water pressure readings recorded in standard units		
	flow_data	https://datamillnorth.org/dataset/2zlgn/yorkshire-water-leakage-dma-15-minute-data	Water flow data recorded by the sensor		Decimal	l/s	Water flow measurements recorded in standardized units		
	temperature		Temperature data recorded by the sensor		Decimal	Celsius	Temperature readings recorded in degrees Celsius (° C) or Fahrenheit (° F)		
asset data	age	https://environment.data.gov.uk/asset-management/drl-app/revision/current/categories/AssetComplex/asset-types/WaterLevelAndFlowMonitoringComplex	Age of the sensor, typically measured from the installation date	The purpose of accessing sensor asset data is to monitor the lifecycle, type, and operational status of sensors, ensuring their reliability for water leakage detection. This data supports proactive maintenance, optimizes sensor deployment, and enhances the overall efficiency of the water monitoring system by minimizing downtime and inaccuracies.	Integer	years	Age of the sensor in years	https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/	5.1, 4.1, 5.7, 5.3, 4.4
	sensor_type	https://sensors-gui.bgs.ac.uk/	Type or classification of the sensor (e.g., flow sensor, pressure sensor)		String	/	Categorical field specifying the type of sensor (e.g., flow sensor, pressure sensor)	MIT License	
	manufacturer	API:https://sensors.bgs.ac.uk/api.html	Company or brand that manufactured the sensor				Name of the sensor's manufacturing company, standardized to the full legal name for consistency		
	installation_year	github:https://github.com/BritishGeologicalSurvey/sensor-things-api-demo/blob/main/sensor-things-api-demo.ipynb	Year when the sensor was installed		Integer	/	Year of sensor installation in four-digit format (YYYY)		
	status		Current operational state of the sensor, indicating whether it is active, inactive, under maintenance, or malfunctioning		String	/	Categorical field representing the sensor’s status, using standardized labels such as 'active,' 'inactive,' 'maintenance,' or 'faulty' for consistency across records and enabling uniform monitoring of sensor functionality		
	model		Model name of the sensor		String	/	Standardized to distinguish between different sensor models and versions		

Pipe Data									
pipe location	pipe_id		Unique indenfier	The purpose of accessing pipe location data is to visualize and analyze the geographical layout of pipelines, which is crucial for pinpointing potential leakage locations. This data supports efficient pipeline maintenance, facilitates rapid response to detected leaks, and enhances the overall management of the water distribution network.	String	/	Unique alphanumeric identifier for each pipeline segment		
	geometry	https://www-nature-com.libproxy.ucl.ac.uk/articles/s41598-024-60840-x/tables/3	Coordinates representing the geographical layout of the pipeline, displayed as a line geometry		Spatial Object	/	Spatial data representing the geographical layout of the pipeline as a line geometry		
asset data	Pipe length		The length of the pipe in meters(m)	The purpose of accessing pipe asset data is to evaluate the physical characteristics, operational status, and historical maintenance of pipelines, which are critical for assessing vulnerability to leaks and failures. This data supports predictive maintenance, enhances the detection of high-risk areas, and improves the overall management of water distribution infrastructure to reduce leakage and service disruptions.	float/decimal	m	Numerical value representing the length of the pipe segment in meters (m)		5.1, 4.1
	Pipe material	https://data-downloads.slip.wa.gov.au/WCORP-002/Shapefile	The material of the pipe section, categorized as Numerical type		String	/	Categorical field describing the material composition of the pipe (e.g., PVC, steel, copper) following industry standard classifications for material types	https://datawa-prod-storage.s3.ap-southeast-2.amazonaws.com/resources/8cefd9c8-79b3-4718-a323-63185023efe0/water-corporation-spatial-data-licence.pdf?Content-Type=application%2Fpdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAZVXRTAT3ITX5SZUN%2F20241209%2Fap-southeast-2%2Fs3%2Faws4_request&X-Amz-Date=20241209T134818Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=5fb7f2ff2219be2abef1a5c6feb1992b75f10d285d69d4c2ceb944f6ce985d77	
	Pipe diameter	data description:https://catalogue.data.wa.gov.au/dataset/water-pipe-wcorp-002	The diameter of pipe in millimeters		float/decimal	cm	Numeric value indicating the diameter of the pipe		
	Status	attribute:https://public-services.slip.wa.gov.au/public/rest/services/SLIP_Public_Services/Infrastructure_and_Utilities/MapServer/20	Current operational status of the pipe (e.g., active, inactive, abandon)		String	/	Categorical field indicating the operational status of the pipe (e.g., active, inactive, abandoned)		
	Year pipe laid		The Installation Year of pipe		Integer	/	Year in four-digit format (YYYY) representing when the pipe was installed		
	Number of breaks		The number of total damages recorded on the pipe		Integer	/	Integer representing the total count of recorded damages or breaks along the pipe segment		
	flow direction		Direction of water flow within the pipe(one-way or two-way).		String	/	Categorical field indicating the direction of water flow within the pipe (e.g., one-way or two-way)		
Repair Data									
repair_id		https://openrepair.org/open-data/downloads/	Unique identifier	The purpose of accessing repair data is to track and analyze the status, location, and timeline of repair activities, helping to address water leakage issues efficiently. This data enables targeted maintenance, reduces downtime, and supports proactive management to	Integer	/	Unique identifier for each repair event	https://creativecommons.org/licenses/by-sa/4.0/	not identified
repair status			Current status of the repair		String/Integer	/	Categorical field representing the current repair status (e.g., pending, in progress, completed)		
reported date/time			Date and time when the repair issue was reported		Datetime	/	Date and time when the repair issue was reported, formatted in ISO 8601 (YYYY-MM-DD HH:MM) for consistent temporal referencing		
Repair start date			Date when the repair work started		Datetime	/	Date when repair work started/completed, in the standard date format (YYYY-MM-DD)		
Repair end date			Date when the repair work was completed		Datetime	/			

Repair location		Location where the repair is being conducted	enhance the resilience and reliability of the water distribution network.	Coordinates	/	Geographical coordinates (e.g., latitude and longitude) representing the exact location of the repair		
Repair type		Type or category of repair required		String/Integer	/	Categorical field specifying the type of repair (e.g., pipe repair, valve replacement) based on a standardized codelist for repair types		
Last maintenance date		Date of the most recent maintenance activity before the repair.		Datetime	/	Date of the last maintenance performed before the repair, recorded in YYYY-MM-DD format		
Fault type		Type of fault or issue that required repair		String/Integer	/	Categorical field indicating the type of fault or issue (e.g., leakage, corrosion, blockage)		